|  |  |
| --- | --- |
|  | Moving Picture, Audio and Data Coding  by Artificial Intelligence  www.mpai.community |

|  |  |
| --- | --- |
| **N829** | 2022/08/24 |
| **Source** | CAE-DC, MMC-DC, and Requirements (ARA) |
| **Title** | Clarifications about Data Formats in MPAI-MMC V2 Call for Technologies |
| **Target** | MPAI-23 |

# Abstract

This document provides clarifications on but does not replace [10]. Those intending to respond to [9] are encouraged to read this document.

# Table of Contents

[Abstract 1](#_Toc111820834)

[Table of Contents 1](#_Toc111820835)

[1 Introduction 2](#_Toc111820836)

[2 General Reference Model 3](#_Toc111820837)

[3 Terms and Definitions 4](#_Toc111820838)

[4 References 5](#_Toc111820839)

[5 Data Formats Functional Requirements 6](#_Toc111820840)

[5.1 Environment Capture 7](#_Toc111820841)

[5.1.1 Audio 7](#_Toc111820842)

[5.1.1.1 Capture Format 7](#_Toc111820843)

[5.1.1.2 Scene description 7](#_Toc111820844)

[5.1.2 Visual 7](#_Toc111820845)

[5.1.2.1 Capture Format 7](#_Toc111820846)

[5.1.2.2 Scene Description 8](#_Toc111820847)

[5.1.3 Combined 8](#_Toc111820848)

[5.1.3.1 Scene Description 8](#_Toc111820849)

[5.1.4 Data 8](#_Toc111820850)

[5.1.4.1 Text 8](#_Toc111820851)

[5.1.4.2 Point of View 8](#_Toc111820852)

[5.2 Feature Extraction 9](#_Toc111820853)

[5.2.1 Text 9](#_Toc111820854)

[5.2.1.1 Text Descriptors 9](#_Toc111820855)

[5.2.2 Speech 9](#_Toc111820856)

[5.2.2.1 Speech Descriptors 9](#_Toc111820857)

[5.2.3 Visual 9](#_Toc111820858)

[5.2.3.1 Face Descriptors 9](#_Toc111820859)

[5.2.3.2 Head Descriptors 10](#_Toc111820860)

[5.2.3.3 Arm Descriptors 10](#_Toc111820861)

[5.2.3.4 Hand Descriptors 10](#_Toc111820862)

[5.2.3.5 Finger Descriptors 11](#_Toc111820863)

[5.2.3.6 Gesture Descriptors 11](#_Toc111820864)

[5.2.4 Combined 12](#_Toc111820865)

[5.2.4.1 Audio-Visual Scene Descriptors 12](#_Toc111820866)

[5.3 Feature Interpretation 12](#_Toc111820867)

[5.3.1 Personal Status 12](#_Toc111820868)

[5.3.1.1 Emotion 12](#_Toc111820869)

[5.3.1.2 Cognitive State 12](#_Toc111820870)

[5.3.1.3 Personal Attitude 13](#_Toc111820871)

[5.3.2 Text 13](#_Toc111820872)

[5.3.2.1 Recognised Text 13](#_Toc111820873)

[5.3.2.2 Personal Status-Text 13](#_Toc111820874)

[5.3.3 Speech 14](#_Toc111820875)

[5.3.3.1 Personal Status-Speech 14](#_Toc111820876)

[5.3.4 Visual 14](#_Toc111820877)

[5.3.4.1 Personal Status-Face 14](#_Toc111820878)

[5.3.4.2 Personal Status-Gesture 14](#_Toc111820879)

[5.3.4.3 Object Identification 15](#_Toc111820880)

[5.3.5 Combined 15](#_Toc111820881)

[5.3.5.1 Personal Status 15](#_Toc111820882)

[5.3.6 Meaning 15](#_Toc111820883)

[5.3.6.1 Meaning 15](#_Toc111820884)

[5.4 Action Generation 16](#_Toc111820885)

[5.5 Experience Generation 16](#_Toc111820886)

[5.5.1 Text 16](#_Toc111820887)

[5.5.1.1 Summary 16](#_Toc111820888)

[5.5.2 Visual 17](#_Toc111820889)

[5.5.2.1 Avatar Model 17](#_Toc111820890)

[5.5.2.2 Avatar Descriptors 17](#_Toc111820891)

[5.5.3 Combined 17](#_Toc111820892)

[5.5.3.1 Environment Model 17](#_Toc111820893)

[5.5.4 Object Metadata 18](#_Toc111820894)

[5.5.4.1 Spatial Attitude 18](#_Toc111820895)

[5.6 Experience Rendering 18](#_Toc111820896)

[5.6.1 Rendered Visual Scene 18](#_Toc111820897)

[Annex 1 - The MPAI AI Framework Model 19](#_Toc111820898)

[Annex 2 - Personal Status Extraction and Display 20](#_Toc111820899)

[Annex 3 - MPAI-MMC V2 Use Cases 22](#_Toc111820900)

[1 Conversation About a Scene 22](#_Toc111820901)

[2 Human-CAV Interaction AIW (HCI) 22](#_Toc111820902)

[3 Avatar-Based Videoconference 24](#_Toc111820903)

# Introduction

In 2021/Q3, the Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) Standards Developing Organisation has published 5 technical specifications covering some of its mission areas:

1. *Governance of the MPAI Ecosystem (MPAI-GME)* [1] lays down the rules that govern ecosystem triggered by MPAI standards.
2. *AI Framework (MPAI-AIF)* [2] specifies an environment (AIF) executing AI Workflows (AIW) composed of AI Modules (AIM).
3. *Multimodal Conversation (MPAI-MMC)* [3] specifies technologies for new forms of human-machine conversation that emulate human-human conversation in completeness and intensity.
4. *Content-based Audio Enhancement (MPAI-CAE)* [4] specifies technologies improving context-based (e.g., in the car, etc.) audio applications (e.g., entertainment, etc.) user experience.
5. *Compression and Understanding of Industrial Data (MPAI-CUI)* [5] specifies technologies to predict the performance of a Company in a prediction horizon, e.g., the probability to default.

In July 2022, MPAI has published Calls for Technologies and associated Functional and Commercial Requirements for 3 new standards:

1. *AI Framework Version 2* [6,7,8].
2. *Multimodal Conversation Version 2* [9,10,11].
3. *Neural Network Watermarking Version 1* [12,13,14].

This document is a revised version of [10] offering a way to navigate and position the technologies called for by [9]. As such, it may be of use to those who intend to respond to the Multimodal Conversation Version 2 (MPAI-MMC V2) Call for Technologies [9].

Disclaimer: This document does NOT supersede [10] which remains the official document providing the functional requirements that technologies proposed in response to [9] should satisfy.

The MPAI-MMC V2 Use Cases assume use of the following technologies:

1. The *AI Workflows* and *AI Modules* specified in MPAI-AIF [2], a summary introduction of which is given by Annex 1 - The MPAI AI Framework Model.
2. The notion of *Personal Status*, i.e., a set of Factors, i.e., internal characteristics of a person or avatar. Currently considered Factors are: Emotion, Cognitive State, and Attitude. Factors are conveyed by 4 Modalities: Text, Speech, Face, and Gesture. Other Modalities are possible but currently not considered.
3. Two Composite AI Modules: *Personal Status Extraction* (PSE) extracting the combined Personal Status of the 4 Modalities and *Personal Status Display* (PSD) animating an avatar based on the Text and the Personal Statuses of the 3 Factors. Annex 2 - Personal Status Extraction and Display gives a summary of the two Composite AIM.

# General Reference Model

Figure 1 is the general MPAI-MMC V2 Reference Model applicable to all MPAI-MMC V2 Use Cases (and to several other MPAI standards as well). A summary description of the Use Cases can be found in Annex 3 - MPAI-MMC V2 Use Cases. The full description is found in [10].

The yellow part in Figure 1 refers to the capture, processing and understanding of the Real Environment, while the blue part refers to the machine’s reaction that is materialised as an experience rendered to the Real Environment.

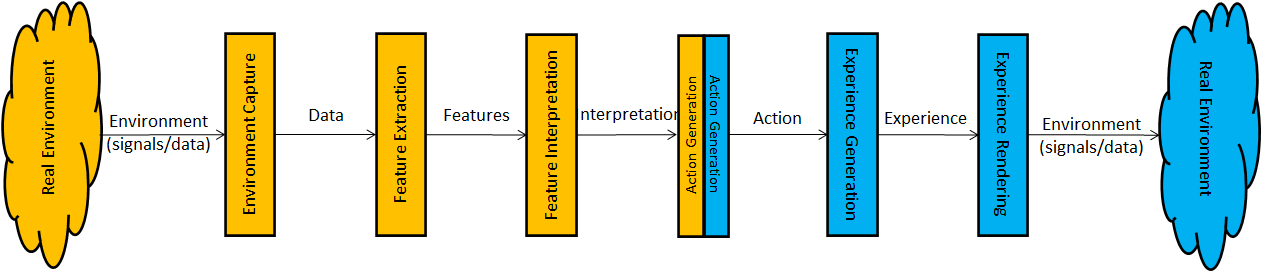


Figure 1 - Reference Model of MPAI-MMC V2

Table 1 describes the functions of each block in Figure 1 called Component in the following.

Table 1 - The Components in the MPAI-MMC V2 Reference Model

|  |  |  |
| --- | --- | --- |
| **#** | **Component** | **Function** |
| 1 | Environment Capture | Captures selected signals (e.g., audio and video) and data (text and other data, e.g., the Point of View of a scene)) from a Real Environment. Typically, the machine creates an internal object-based representation of the captured data. |
| 2 | Feature Extraction | Extracts Features of interest from data. Features are embedded in an Object allowing its classification as an Object of a particular type. |
| 3 | Feature Interpretation | Derives semantics from Features. |
| 4 | Response Generation | Generates a congruent Action from Interpretations. |
| 5 | Experience Generation | Generates an Experience from the Action. The Experience can be the synthetic speech version of the text and a properly animated avatar face. |
| 6 | Environment Delivery | Renders the Experience to a target Real Environment in a way that can be perceived by a human. |

Figure 2 extends Figure 1 by adding the data formats handled by each Component of the MPAI-MMC V2 Reference Model.

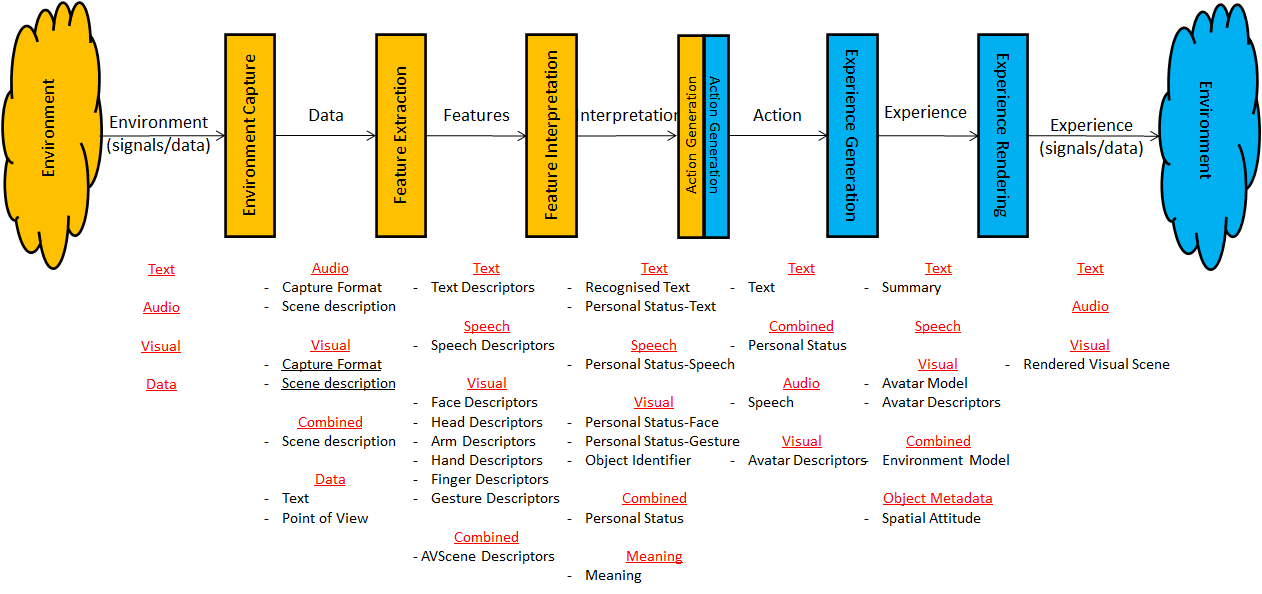


Figure 2 - Data Formats of the MPAI-MMC V2 Reference Model

# Terms and Definitions

Table 2 defines the Terms used in this document.

Table 2 - Terms and Definitions of MPAI-MMC V2

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Captured Data | The data acquired either by converting analogue signals into data or by directly acquiring data from a Real Environment. |
| Environment | A portion of a Real or Virtual Space. |
| Experience | The content generated by a machine as a result of its interaction with a human or an avatar. |
| Extraction | The process of extracting Features from data. |
| Factor | One of Emotion, Cognitive State or Personal Attitude. |
| Feature | Elements embedded in an Object indicating that it should be classified as an Object of a particular type. |
| Gesture | The ensemble of the head, arm, hand, and finger movements of a Human Object or an Avatar. |
| Interpretation | The process of deriving semantics from Features. |
| Meaning | Information extracted from Text, Personal Status, and other elements, e.g., Object ID, such as question, statement, exclamation, expression of doubt, request, invitation. |
| Modality | One of Text, Speech, Face, or Gesture. |
| Object | Data that is perceived as an object or a combination of objects when rendered. |
| * Arm | The Visual Object extracted from a Human Object that is perceived as an arm when rendered. |
| * Audio | Data that is perceived by the ears when rendered. |
| * Face | The Visual Object extracted from a Human Object that is perceived as a face when rendered. |
| * Finger | The Visual Object extracted from a Human Object that is perceived as a finger when rendered. |
| * Hand | The Visual Object extracted from a Human Object that is perceived as a hand when rendered. |
| * Head | The Visual Object extracted from a Human Object that is perceived as a head when rendered. |
| * Human | The Visual Object extracted from a Human Object that is perceived as a human when rendered. |
| * Visual | Data that is perceived by the eyes when rendered. |
| Orientation | The 3 yaw, pitch, and roll (α,β,γ) angles of a representative point of an object in the Real and Virtual Space. |
| Personal Status | The ensemble of information internal to a human, including Emotion, Cognitive State, and Attitude. |
| * Cognitive State | A typically rational Factor of the internal status of a human or avatar reflecting the way they understand the Environment, such as “Confused”, “Dubious”, “Convinced”. |
| * Emotion | A typically irrational Factor of the internal status of a human or avatar resulting from their interaction with the Environment or subsets of it, such as “Angry”, “Sad”, “Determined”. |
| * Personal Attitude | A Factor of the internal status of a human or avatar related to the way they intend to position themselves vis-à-vis the Environment or subsets of it, e.g., “Respectful”, “Confrontational”, “Soothing”. |
| Phonetic Unit | One of word, syllable, or phoneme. |
| Point of View | The Spatial Attitude of a human or avatar looking at an Environment. |
| Position | The 3 coordinates (x,y,z) of a representative point of an object in the Real and Virtual Space. |
| Rendered Scene | The output produced when rendering an Audio-Visual Scene from a selected Point of View. |
| Scene | An Environment populated by humans and real objects (in a Real Environment) or by avatars and virtual objects (in a Virtual Environment). |
| Spatial Attitude | The set of 18 values (x,y,z,α,β,γ; their 1st order derivatives; and their 2nd order derivatives) corresponding to the Position (x,y,z) and Orientation (α,β,γ) and their derivatives of an Object. The position of the object corresponds to that of a representative point of the object, if the object is rigid. |

# References

This document references the following documents:

1. MPAI; Technical Specification: The Governance of the MPAI Ecosystem V1: <https://mpai.community/standards/mpai-gme/>.
2. MPAI; Technical Specification: AI Framework (MPAI-AIF) V1.1; <https://mpai.community/standards/mpai-aif/>.
3. MPAI; Technical Specification: Specification: Multimodal Conversation (MPAI-MMC) V1.2; <https://mpai.community/standards/mpai-mmc/>.
4. MPAI; Technical Specification: Context-based Audio Enhancement (MPAI-CAE) V1.4; [https://mpai.community/standards/resources/#CAE](https://mpai.community/standards/resources/" \l "CAE)
5. MPAI; Technical Specification: Compression and Understanding of Industrial Data (MPAI-CUI) V1.1; <https://mpai.community/standards/resources/#CUI>
6. MPAI-AIF V2 Call for Technologies; <https://mpai.community/standards/mpai-aif/call-for-technologies/mpai-aif-v2-call-for-technologies/>.
7. MPAI-AIF V2 Use Cases and Functional Requirements; <https://mpai.community/standards/mpai-aif/use-cases-and-functional-requirements/mpai-aif-v2-use-cases-and-functional-requirements/>.
8. MPAI-AIF V2 Framework Licence; <https://mpai.community/standards/mpai-aif/framework-licence/mpai-aif-v2-framework-licence/>
9. MPAI; MPAI-MMC V2 Call for Technologies; <https://mpai.community/standards/mpai-mmc/call-for-technologies-2/mpai-mmc-v2-call-for-technologies/>
10. Multimodal Conversation (MPAI-MMC) V2 Use Cases and Functional Requirements; <https://mpai.community/standards/mpai-mmc/use-cases-and-functional-requirements/mpai-mmc-v2-use-cases-and-functional-requirements/>
11. MPAI; MPAI-MMC V2 Framework Licence; <https://mpai.community/standards/mpai-mmc/framework-licence/mpai-mmc-v2-framework-licence/>
12. MPAI; MPAI-NNW V2 Call for Technologies; <https://mpai.community/standards/mpai-mmc/call-for-technologies/mpai-nnw-call-for-technologies/>
13. Neural Network Watermarking (MPAI-NNW) Use Cases and Functional Requirements; MPAI N793; <https://mpai.community/standards/mpai-nnw/use-cases-and-functional-requirements/>
14. Neural Network Watermarking (MPAI-NNW) Framework Licence;; <https://mpai.community/standards/mpai-nnw/framework-licence/>
15. Universal Coded Character Set (UCS): ISO/IEC 10646; December 2020

# Data Formats Functional Requirements

This chapter is subdivided in the 6 Components of the Reference Model. For each Component it identifies the major data types involved – e.g., Audio, Visual etc. – and sub-data types – e.g., Capture Format and Scene Description. For each the following is provided

|  |  |
| --- | --- |
| Context | Where the data format is used. |
| Goal | What is the purpose of the data format. |
| Requirements | What are the requirements of the data format . |
| Comments | Any comment. |
| Requested | What is requested from respondents. |

## Environment Capture

### Audio

#### Capture Format

|  |  |
| --- | --- |
| Context | In the CAV-HCI and MCS-ABV Use Cases, a machine captures speech from a noisy scene. |
| Goal | To produce a data stream that preserves the information required by subsequent processing. |
| Requirements | The representation format should enable the production of the Description of the Audio Scene. |
| Comments | MPAI has standardised Interleaved Multichannel Audio in [4] which is believed to be adequate for the task. However, alternative representation formats will be considered. |
| Requested | The output format of a microphone array satisfying the Requirements. |

#### Scene description

|  |  |
| --- | --- |
| Context | The CAV-HCI and MCS-ABV require technology to Describe an Audio Scene, where Audio is composed of Speech Objects in a noisy environment. |
| Goal | To accurately represent the Speech Objects of the audio scene with their positions. Speech objects can change their position and orientation. If the sound source is spherical, orientation information is not required. |
| Requirements | The Audio Scene Description shall provide:   1. Access to the individual Speech Objects. 2. The spatial coordinates and the orientation of the individual objects in the coordinate system. |
| Comments | The Audio Scene Description standardised in [4] is considered adequate to the MCS-ABV Use Case and the CAV-HCI Use Case in the in-cabin sub-use case. Extensions are believed to be required to support the CAV-HCI in the outdoor sub-use case. |
| Requested | The Audio (Speech) Scene Description satisfying the Requirements. |

### Visual

#### Capture Format

|  |  |
| --- | --- |
| Context | In the CAV-HCI, MCS-ABV and MCS-CAS Use Cases, a machine captures a visual scene. |
| Goal | To produce a data stream that preserves the information required by a separate AIM to produce the Visual Scene Description. |
| Requirements | The representation format shall support the following video capture formats:   1. 2D 2. 2D+depth 3. 3D   and enable the production of the Description of the Visual Scene. |
| Comments |  |
| Requested | The output format of a device capturing the visual scene satisfying the Requirements. |

#### Scene Description

|  |  |
| --- | --- |
| Context | In the CAV-HCI, MCS-ABV and MCS-CAS Use Cases, a machine captures a visual scene and produces a Visual Scene Description. |
| Goal | To produce an accurate representation of the visual objects of the scene, assumed to be static, with their position and orientation. |
| Requirements | The format shall enable the representation of Human Objects with the ability to extract components such as face and head. |
| Comments |  |
| Requested | The Visual Scene Description satisfying the Requirements. |

### Combined

#### Scene Description

|  |  |
| --- | --- |
| Context | In the CAV-HCI and MCS-ABV Use Cases, a machine captures the audio-visual scene and produces an Audio-Visual Scene Description. |
| Goal | To provide an accurate representation of the scene’s Audio-Visual Objects assumed to be static. |
| Requirements | The Audio-Visual Scene Description shall be able to represent the association of Speech Objects with Visual Objects. |
| Comments |  |
| Requested | The Audio-Visual Scene Description satisfying the Requirements. |

### Data

#### Text

|  |  |
| --- | --- |
| Context | The CAV-HCI and MCS-ABV Use Cases contemplate the possibility for users to communicate with a machine using Text. |
| Goal | To represent the characters of alphabets. |
| Requirements | The format shall represent the characters currently most used worldwide. |
| Comments | MPAI-MMC V1 [3] has adopted the Universal Coded Character Set (UCS). The use of UCS has been extended to all cases where character sets are needed. |
| Requested | Universal Coded Character Set (UCS) is considered adequate. MPAI is willing to reconsider its choice if motivated arguments are provided. |

#### Point of View

|  |  |
| --- | --- |
| Context | In the MCS-ABV Use Case, a human taking part in the virtual videoconference sets the coordinates of the point and the orientation from which they will view the virtual room. |
| Goal | To define the format of the Point of View. |
| Requirements | The format shall be able to represent Spatial Attitude of the Point of View and shall also be able to represent the 1st and 2nd order derivatives (velocity and acceleration). |
| Comments |  |
| Requested | A point of View format satisfying the Requirements. |

## Feature Extraction

### Text

#### Text Descriptors

|  |  |
| --- | --- |
| Context | All Use Cases use Text either as human- or machine-generated information. |
| Goal | To define Feature that will allow the Feature Interpretation to extract the Personal Status from Text. |
| Requirements | The Text Descriptors shall enable a subsequent Feature Interpretation AIM to extract the Personal Status of Text. |
| Comments | MPAI-MMC V1 [3] has adopted Universal Coded Character Set (UCS). |
| Requested | Text Descriptors satisfying the Requirements. |

### Speech

#### Speech Descriptors

|  |  |
| --- | --- |
| Context | In all MPAI-MMC V2 Use Cases, there is a need to extract Speech Features and represent them as Descriptors. |
| Goal | To enable   1. Speech-based identification of a Speech Object as produced by a Human Object representing a human belonging to a group composed of a limited number of humans. 2. Extraction of the Personal Status of a Speech Object. |
| Requirements | The Descriptors shall enable a party to achieve any of the Speech Descriptors’ Goals using the Speech Descriptors generated by another party. |
| Comments | Respondents may propose a solution for a subset or for all Goals. |
| Requested | Speech Descriptors satisfying the Requirements. |

### Visual

#### Face Descriptors

|  |  |
| --- | --- |
| Context | All MPAI-MMC V2 Use Cases need Face Features represented as Descriptors. Other usages of the Features are also possible. |
| Goal | To describe a Face Object for the purpose of:   1. Identifying an instance of a Face Object as belonging to a Human Object representing a human belonging to a group composed of a limited number of humans. 2. Extracting the Personal Status of a Face Object. 3. Representing the Face component of the Avatar Model having a face as the Face Object. 4. Animating an Avatar Model. 5. Tracking the Spatial Attitude of the point representing a moving Face Object considered as a rigid body. |
| Requirements | The Descriptors shall enable a party to achieve any of the Face Descriptors’ Goals using the Face Descriptors generated by another party. |
| Comments | It is desirable that the Descriptors of the face of an Avatar be the same as the Descriptors of a Face Object. However, this is not a requirement.  Respondents may propose solutions for a subset or for all Goals. |
| Requested | Face Descriptors satisfying the Requirements. |

#### Head Descriptors

|  |  |
| --- | --- |
| Context | MPAI-MMC V2 Use Cases where Personal Status is used need to extract Head Features represented as Descriptors. Other usages of the Features are also possible. |
| Goal | To describe a Head Object for the purpose of:   1. Identifying an instance of a Head Object as belonging to a Human Object representing a human belonging to a group composed of a limited number of humans. 2. Extracting the Personal Status of the Gesture that depends on the Head. 3. Representing the Head component of the Avatar Model having a head as the Head Object. 4. Animating the Head of an Avatar Model. 5. Tracking the Spatial Attitude of the point representing the Head considered as a rigid body. 6. Describing the trajectory of the point representing a moving Head Object. |
| Requirements | The Descriptors shall enable a party to achieve any of the Head Descriptors Goals using the Head Descriptors generated by another party. |
| Comments | Respondents may propose solutions for a subset or for all Goals. |
| Requested | Head Descriptors satisfying the Requirements. |

#### Arm Descriptors

|  |  |
| --- | --- |
| Context | In all MPAI-MMC V2 Use Cases where Personal Status is used, there is a need to extract Arm Features represented as Descriptors. Other usages of the Features are also possible. |
| Goal | To describe an Arm Object for the purpose of:   1. Extracting the Personal Status of the Gesture that depends on the Arm. 2. Representing the Arm component of the Avatar Model having an arm as the Arm Object. 3. Animating the Arm of an Avatar Model. 4. Tracking the Spatial Attitude of the points representing the Arm considered as two connected rigid bodies. 5. Describing the trajectory of the points representing a moving Arm Object. |
| Requirements | The Descriptors shall enable a party to achieve any of the Arm Descriptors Goals using the Arm Descriptors generated by another party. |
| Comments | Respondents may propose solutions for a subset or for all Goals. |
| Requested | Arm Descriptors satisfying the Requirements. |

#### Hand Descriptors

|  |  |
| --- | --- |
| Context | In all MPAI-MMC V2 Use Cases where Personal Status is used, there is a need to extract Hand Features represented as Descriptors. Other usages of the Features are also possible. |
| Goal | To describe a Hand Object for the purpose of:   1. Extracting the Personal Status of the Gesture that depends on the Hand. 2. Representing the Hand component of the Avatar Model having a hand as the Hand Object. 3. Animating the Hand of an Avatar Model. 4. Tracking the Spatial Attitude of the point representing the Hand considered as a rigid body. 5. Describing the trajectory of the point representing a moving Hand Object. |
| Requirements | The Descriptors shall enable a party to achieve any of the Hand Descriptors Goals using the Hand Descriptors generated by another party. |
| Comments | Respondents may propose solutions for a subset or for all Goals. |
| Requested | Hand Descriptors satisfying the Requirements. |

#### Finger Descriptors

|  |  |
| --- | --- |
| Context | In all MPAI-MMC V2 Use Cases where Personal Status is used, there is a need to extract Finger Features represented as Descriptors. Other usages of the Features are also possible. |
| Goal | To describe a Finger Object for the purpose of:   1. Extracting the Personal Status of the Gesture that depends on the Finger. 2. Representing the Finger component of the Avatar Model having a finger as the Finger Object. 3. Animating the Finger of an Avatar Model. 4. Tracking the Spatial Attitude of the points representing the Finger considered as a rigid body. 5. Describing the trajectory of the point representing a moving Finger Object. |
| Requirements | The Descriptors shall enable a party to achieve any of the Finger Descriptors’ Goals using the Finger Descriptors generated by another party. |
| Comments | Respondents may propose solutions for a subset or for all Goals. |
| Requested | Finger Descriptors satisfying the Requirements. |

#### Gesture Descriptors

|  |  |
| --- | --- |
| Context | In all MPAI-MMC V2 Use Cases where Personal Status is used, there is a need to extract Gesture Features represented as Descriptors. Other usages of the Features are also possible. |
| Goal | To provide an organised composition of Head Descriptors, Arm Descriptors, Hand Descriptors, and Finger Descriptors to:   1. Represent arbitrary movement of head, arms, hands, and fingers. 2. Recognise:    1. Sign language.    2. Culture-dependent signs (e.g., mudra sign).    3. Coded hand signs, e.g., to indicate a particular object in a scene.    4. A human’s Personal Status conveyed by Gesture. 3. Animate an avatar’s head, arms, hands, and fingers. |
| Requirements | Gesture Descriptors shall be a combination of Head, Arm, Hand, and Finger Descriptors defined in this document. |
| Request | Syntax and semantics of head, arm, hand, and finger key points that are independent of the technology used to implement the Key Point Detection AIM – i.e., image processing, or ML, or combinations of the two. |
| Comments | Respondents may propose solutions for a subset or for all Goals. |
| Requested | Gesture Descriptors satisfying the Requirements. |

### Combined

#### Audio-Visual Scene Descriptors

|  |  |
| --- | --- |
| Context | All MPAI-MMC V2 Use Cases require a standard representation of audio-visual scenes. |
| Goal | To describe an audio-visual 3D space (a room, a square, a mall, etc.) made of separate audio, video and audio-visual objects. |
| Requirements | The Descriptors shall enable a party to synthesise a virtual replica of an audio-visual 3D space by simply using the corresponding Descriptors. |
| Comments |  |
| Requested | Audio-Visual Scene Descriptors satisfying the Requirements. |

## Feature Interpretation

### Personal Status

#### Emotion

|  |  |
| --- | --- |
| Context | All Use Cases using the Personal Status Display Composite AIM require a standard representation of Emotion. |
| Goal | To provide a representation of Emotion. |
| Requirements | The same format shall be usable in all current Modalities with possible Modality-specific extensions. |
| Comments | In [3] MPAI has standardised a set of basic Emotions and their semantics. However, alternative representations of Emotion will be considered if they bring demonstrated superior performance. |
| Requested | Representation of Emotion satisfying the Requirements. |

#### Cognitive State

|  |  |
| --- | --- |
| Context | All Use Cases using the Personal Status Display Composite AIM require a standard representation of Cognitive State. |
| Goal | To provide a representation of Cognitive State. |
| Requirements | The same format shall be usable in all current Modalities with possible Modality-specific extensions. |
| Comments | In [3] MPAI has standardised a set of basic Emotions and their semantics. A representation of Cognitive State based on the same principles is a desirable feature. However, representations of Cognitive State based on different principles will be considered if they bring demonstrated superior performance. |
| Requested | Representation of Cognitive State satisfying the Requirements. |

#### Personal Attitude

|  |  |
| --- | --- |
| Context | All Use Cases using the Personal Status Display Composite AIM require a standard representation of Personal Attitude. |
| Goal | To provide a representation of Personal Attitude. |
| Requirements | The same format shall be usable in all current Modalities with possible Modality-specific extensions. |
| Comments | In [3] MPAI has standardised a set of basic Emotions and their semantics. A representation of Personal Attitude based on the same principles is a desirable feature. However, representations of Personal Attitude based on different principles will be considered if they bring demonstrated superior performance. |
| Requested | Representation of Personal Attitude satisfying the Requirements. |

### Text

#### Recognised Text

|  |  |
| --- | --- |
| Context | An AIM may need information about the hypotheses made by a Speech Recogniser. |
| Goal | To provide information on the hypotheses considered by an Automatic Speech Recogniser before outputting the final text. |
| Requirements | The data format shall provide:   1. Reference time: absolute time or the time the Speech Recognition process has started operation, or the time the Phonetic Unit starts. 2. The Phonetic Unit type about which information is provided. 3. Starting time (when the start time is not measured from the start of the Phonetic Unit) and duration of the Phonetic Unit. 4. Hypotheses relative to the Phonetic Unit and the corresponding probabilities. |
| Comments | A candidate solution (no endorsement) is kaldi (https://kaldi-asr.org/). |
| Requested | Representation of Recognised Text satisfying the Requirements. |

#### Personal Status-Text

|  |  |
| --- | --- |
| Context | Most MPAI-MMC V2 Use Cases using Personal Status need to extract Personal Status-Text. |
| Goal | To represent the Personal Status extracted from Text. |
| Requirements | The Personal Status extracted from Text shall include:   1. At least one Factor. 2. Selected Descriptors used to extract Personal Status-Text (optional). 3. The Phonetic Unit from which the Personal Status-Text starts. |
| Comments |  |
| Requested | Representation of Personal Status-Text satisfying the Requirements. |

### Speech

#### Personal Status-Speech

|  |  |
| --- | --- |
| Context | Most MPAI-MMC V2 Use Cases using Personal Status need to extract Personal Status-Speech. |
| Goal | To represent the Personal Status extracted from a Speech Object. |
| Requirements | The Personal Status extracted from Speech shall include:   1. At least one Factor. 2. Selected Descriptors used to extract Personal Status-Speech (optional). 3. The start time of Personal Status-Speech. 4. Indication of time measure:    1. The start time of Personal Status Extraction operation.    2. Absolute time |
| Comments |  |
| Requested | Representation of Personal Status-Speech satisfying the Requirements. |

### Visual

#### Personal Status-Face

|  |  |
| --- | --- |
| Context | Most MPAI-MMC V2 Use Cases using Personal Status need to extract Personal Status-Face. |
| Goal | To represent the Personal Status extracted from a Face Object. |
| Requirements | The Personal Status extracted from Face shall include:   1. At least one Factor. 2. Selected Descriptors used to extract Personal Status-Face (optional). 3. The start time of Personal Status-Face. 4. Indication of time measure:    1. The start time of Personal Status Extraction operation.    2. Absolute time |
| Comments |  |
| Requested | Representation of Personal Status-Face satisfying the Requirements. |

#### Personal Status-Gesture

|  |  |
| --- | --- |
| Context | Most MPAI-MMC V2 Use Cases using Personal Status need to extract Personal Status-Gesture. |
| Goal | To represent the Personal Status extracted from Gesture. |
| Requirements | The Personal Status extracted from Gesture shall include:   1. At least one Factor. 2. Selected Descriptors used to extract Personal Status-Gesture (optional). 3. The start time of Personal Status-Gesture. 4. Indication of time measure:    1. The start time of Personal Status Extraction operation.    2. Absolute time |
| Comments |  |
| Requested | Representation of Personal Status-Gesture satisfying the Requirements. |

#### Object Identification

|  |  |
| --- | --- |
| Context | In CAV-HCI and MCS-CAS, there is a need to represent the ID of a Generic Object. |
| Goal | To provide Identification of a Generic Object. |
| Requirements | The Object Identifier shall identify individual objects. |
| Comments | MPAI-MMC V1 [3] already specifies an Identifier for the identification of an object. |
| Requested | Object Identification format satisfying the Requirements. |

### Combined

#### Personal Status

|  |  |
| --- | --- |
| Context | In all MPAI-MMC V2 Use Cases, the Personal Status is used as a data structure exchanged between AIMs. |
| Goal | To represent the time-varying Personal Status:   1. Represented by (a subset of) 3 Factors. 2. Conveyed by (a subset of) 4 Modalities. |
| Requirements | The Personal Status shall satisfy the following functional requirements:   1. All Factors need not be present. 2. All Factors need not be conveyed by all present Modalities. 3. The Factors and Modalities need not be present all the time. 4. A Factor in a Modality may change with time. 5. The fused Personal Status, i.e., the common value of the Factors in all Modalities, if one exists, shall be provided. 6. The differential information from one Personal Status versus the preceding value of the Personal Status shall be efficiently encoded as a change. |
| Comments |  |
| Requested | Representation of Personal Status satisfying the Requirements. |

### Meaning

#### Meaning

|  |  |
| --- | --- |
| Context | In all MPAI-MMC Use Cases, there is a need to extract the Meaning of a Text and MPAI-MMC V1 already specifies its syntax and semantics. |
| Goal | To provide the Meaning of a Text. |
| Requirements | The format shall include an extensible list of meanings and their digital repres­en­tations satisfying the following requirements:   1. The meaning extracted from the input text shall have a structure that includes grammatical information and semantic information. 2. The digital representation of meaning shall allow for the addition of new features to be used in different applications. |
| Comments | MPAI-MMC V1 [3] already specifies syntax and semantics of Meaning. |
| Requested | Representation of Meaning satisfying the Requirements. |

## Action Generation

The 3 Use Cases considered by MPAI-MMC V2 can be divided in two classes:

1. The two Use Cases based on a human-machine conversation (CAV-HCI and MCS-CAS) and the sub-use case where the Virtual Secretary has a conversation with the Avatars (MC-ABV) where the Action Generation Component (Question and Dialogue Processing AIM) develops a response. Table 3 gives the input and output data of the Action Generation Component:

Table 3 - Input/Output data of Question and Dialogue Processing

|  |  |
| --- | --- |
| **Input** | **Output** |
| Language-Understanding Text  Meaning  Personal Status  User Identity | Text  Personal Status |

In a more general case, the Action Generation Component could forward the Spatial Attitude of the machine and of the human the machine is talking to for the Experience Generation Component to be able to generate an avatar that properly gazes at the human it is talking to.

1. The MCS-ABV Use Cases where an Action is produced by
   1. The Avatar Description AIM in the transmitting client integrating:
      1. The Face and Gesture Descriptors.
      2. The Personal Status.
      3. The Meaning.
   2. The Speech Translation AIM in the server translating the speech streams based on the Language Preferences.
   3. The participant-driven receiving client arranging the virtual videoconference room that the participant can navigate.

## Experience Generation

### Text

#### Summary

|  |  |
| --- | --- |
| Context | In the MCS-ABV Use Case, the Virtual Secretary produces a summary by listening to the utterances of the participating avatars and detecting their Personal Statuses. Participating avatars may comment and the Virtual Secretary may revise the summary. |
| Goal | To produce a summary that facilitates the creation of comments by participating avatars and the processing of comments by the Virtual Secretary. |
| Requirements | The Summary shall be usable by humans and by machines for review. |
| Comments | Text with metadata is a candidate format. The metadata can be based on the Personal Status of the avatar whose sentence the Virtual Secretary’s Summary reports. The avatars reviewing the summary can process the metadata of the summary to better understand the Virtual Secretary’s intentions. An avatar whose utterance has not been properly reported in the Summary can propose a replacement with its own version of the Personal Status. |
| Requested | Representation of Summary satisfying the Requirements. |

### Visual

#### Avatar Model

|  |  |
| --- | --- |
| Context | In the MPAI-ABV Use Case, a client receives Avatar Models from the humans participating in the videoconference to be animated with each participant’s Avatar Descriptors. |
| Goal | To provide a standard Representation of a static avatar exposing handles to allow its animation using Descriptors. |
| Requirements | The Avatar Model shall enable a party to synthesise a static avatar which is a faithful reproduction of a real human by simply using the Avatar Model generated by another party. Currently, the avatar Model is required to only cover the upper part of the body (from the waist up). |
| Comments | The ability to represent the rest of the body is desirable but not required. |
| Requested | Representation of Avatar Model satisfying the Requirements. |

#### Avatar Descriptors

|  |  |
| --- | --- |
| Context | In the MPAI-ABV Use Case, a client receives Avatar Models from the humans participating in the videoconference to be animated by each participant’s Avatar Descriptors. |
| Goal | To describe the instantaneous changes of the face, head, arms, hands, and fingers of an Avatar Model. |
| Requirements | The Descriptors shall enable a party to animate an Avatar Model by simply using the Avatar Descriptors generated by another party. |
| Comments |  |
| Requested | Representation of Avatar Descriptors satisfying the Requirements. |

### Combined

#### Environment Model

|  |  |
| --- | --- |
| Context | In the MCS-ABV Use Case, the provider of the Avatar-Based Videoconference Service distributes the Model of the Environment where the virtual videoconference is held. |
| Goal | To describe a *virtual* audio and visual 3D space (a room, a square, a mall, etc.) made of objects with handles allowing an application to move them around. |
| Requirements | The Model shall enable a party to synthesise a virtual audio-visual environment, e.g., a room or a mall or a square, by simply using the standard Environment Model created by another party.  Currently the format should support Environments with the following features:   1. Lighting is uniform. 2. Walls reflect an attenuated sound from an impinging directional sound. |
| Comments |  |
| Requested | Representation of Environment Model satisfying the Requirements. |

### Object Metadata

#### Spatial Attitude

|  |  |
| --- | --- |
| Context | In the MCS-ABV Use Case, the virtual videonference room is populated with avatars with an assigned Spatial Attitude that does not include velocity and acceleration. |
| Goal | To provide a format to express the Spatial Attitude of an Object. |
| Requirements | To represent Spatial Attitude as a set of 18 values (x,y,z,α,β,γ; their 1st order derivatives; their 2nd order derivatives) corresponding to the Position and Orientation and their derivatives of an Object in a Real or Virtual Environment. |
| Comments | In the MPAI-MMC V2 Use Cases, velocity and acceleration are not used. However, MPAI is studying Use Cases of other areas (e.g., Connected Autonomous Vehicles) where representation of velocity and acceleration is required. |
| Requested | Representation of Spatial Attitude satisfying the Requirements. |

## Experience Rendering

The MPAI-MMC V2 Use Cases do not consider the rendering of generated audio-visual signals. It is generically assumed that the user enjoy the experience using a spatial audio and 3D Video device.

### Rendered Visual Scene

|  |  |
| --- | --- |
| Context | In the MCS-CAS Use Case, the human has the possibility of viewing him/herself how the machine sees the scene. |
| Goal | To display the 3D representation of the Visual Scene perceived by the machine. |
| Requirements | The Rendered Visual Scene shall have a format that enables the graphic rendering of the scene. |
| Comments |  |
| Requested | A Rendered Visual Scene satisfying the Requirements. |

1. The MPAI AI Framework Model

The common infrastructure enabling the implementation of MPAI Application Standards and access to the MPAI Store is the AI Framework (AIF) Standard (MPAI-AIF). Figure 1 depicts the MPAI-AIF Reference Model under which Implementations of MPAI Application Standards and user-defined MPAI-AIF conforming applications operate.

An AIF Implementation allows the execution of AI Workflows (AIW), composed of basic processing elements called AI Modules (AIM).

Diagram

Description automatically generated

Figure 3 - The AI Framework (AIF) Reference Model and its Components

MPAI Application Standards normatively specify the Function of the AIW and its AIMs, the Syntax and Semantics of the input and output Data of the AIW and its AIMs, and the Connections between and among the AIMs of an AIW. MPAI Application Standards do not specify the internal architecture of the AIMs, which may be based on AI or data processing technologies, and be implemented in software, hardware or mixed software and hardware technologies.

MPAI defines Interoperability as the ability to replace an AIW or an AIM Implementation with a functionally equivalent Implementation. Three Interoperability Levels of an AIW executed in an AIF are defined:

*Level 1 –* Implementer-specific and conforming with the MPAI-AIF Standard.

*Level 2 –* Conforming with an MPAI Application Standard.

*Level 3 –* Conforming with an MPAI Application Standard and certified by a Performance As­sessor.

MPAI offers Users access to the promised benefits of AI with a guarantee of increased transparency, trust and reliability of Implementations as the Interoperability Level of an Implementation moves from 1 to 3.

1. Personal Status Extraction and Display

Personal Status is defined as a set of internal characteristics of a person or avatar. MPAI currently considers Emotion, Cognitive State, and Attitude. It can be conveyed by the following Modalities: Text, Speech, Face, and Gesture. Other Modalities are possible but currently not considered.

**Personal Status Extraction (PSE)** is a composite AIM that analyses the Manifestation of a Personal Status conveyed by any of the Text, Speech, Face, and Gesture Modalities and provides the estimate of the Personal Status. Note that Personal Status applies to both humans and avatars. This Composite AIM is used in the MPAI-MMC Use Cases as a replacement for the combination of the AIMs depicted in Figure 4.

The estimate of the Personal Status of a human or an avatar is the result of an analysis carried out on each available Modality in 3 steps:

1. *Data Capture* (e.g., characters and words, a digitised speech segment, the digital video containing the hand of a person, etc.).
2. *Feature Extraction* (e.g., pitch and intonation of the speech segment, thumb of the hand raised, the right eye winking, etc.).
3. *Personal Status Interpretation* (i.e., one, two, or all three of Emotion, Cognitive State and Attitude).

Figure 4 depicts the Personal Status estimation process used in the MPAI-MMC V2 Use Cases:

1. Features are extracted from Text, Speech, Face Object, and Human Object and expressed as Descriptors.
2. Descriptors are interpreted and the specific Manifestations of the Personal Status in the Text, Speech, Face, and Gesture channels derived.
3. The different estimates of the Personal Status are fused to the Personal Status.

Diagram

Description automatically generated

Figure 4 – Reference Model of Personal Status Extraction

An implementation can combine, e.g., the PS-Gesture Description and PS-Gesture Interpretation AIMs into one AIM, and directly provide PS-Gesture from a Human Object without exposing PS-Gesture Descriptors.

**The Personal Status Display (PSD)** the Composite AIM depicted in *Figure 5* generates a speaking avatar experience. PSD receives Text and Personal Status from Response Generation and generates an avatar optionally forwarding Text and producing:

1. Synthesised Machine Speech using Text and Personal Status of PS (Speech).
2. Avatar Face using Machine Speech, Text, and PS (Face).
3. Avatar Gesture using PS (Gesture) and Text.

Diagram

Description automatically generated

*Figure 5 – Reference Model of Personal Status Display*

1. MPAI-MMC V2 Use Cases

This Annex provides a general description of the three MPAI-MMC V2. More details ate provided by [10].

# Conversation About a Scene

A *human* holds a conversation with a machine indicating the object in the Environment s/he wishes to talk or ask questions about.

The *machine*:

1. Sees and hears the Environment containing the human and some scattered objects.
2. Understands the human’s Speech and gets the human’s Personal Status.
3. Understands which object the human is referring to.
4. Generates an avatar that:
   1. Utters Speech conveying a synthetic Personal Status that is pertinent with the human’s Personal Status Manifested in his/her Speech, Face, and Gesture.
   2. Shows a face conveying a Personal Status that is pertinent with the human’s Personal Status and the its own Personal Status resulting from the conversation.
5. The machine labels the Visual Objects of its internally developed representation of the Visual Scene and renders it to the human.

The human can use the rendered scene to understand how the machine sees the Environment.

Diagram, schematic

Description automatically generated

Figure 6 – Reference Model of Conversation About a Scene

# Human-CAV Interaction AIW (HCI)

This use case is part of the Connected Autonomous Vehicle (CAV) project. The CAV executes commands to displace itself based on 1) the analysis and interpretation of the data sensed by a range of onboard sensors exploring the environment and 2) the information transmitted by other sources in range, e.g., other CAVs, traffic lights and roadside units.

Figure 7 depicts the four CAV subsystems.

Diagram

Description automatically generated

Figure 7 – The Connected Autonomous Vehicle Reference Model

**Human-CAV interaction (HCI)** recognises the human owner or renter, responds to humans’ commands and queries and senses human activities during the travel and may activate other Subsystems in response to humans’ requests. The data exchanged between HCI and the Autonomous Motion Subsystem (AMS) is depicted in Figure 8 but the requirements of the format of the data exchanged between HCI and AMS are not part of this document.

**Environment Sensing Subsystem (ESS)** acquires information from the Environment via a variety of sensors and produces a representation of the Environment (Basic World Representation), i.e., its best understanding of the Environment based on the sensed data.

**Autonomous Motion Subsystem (AMS)** computes the route to destination, uses different sources of information – CAV sensors, other CAVs and transmitting units – to produce a Full World Representation and issues commands to drive the CAV to the intended destination.

**Motion Actuation Subsystem (MAS)** provides non-electromagnetic and non-acoustical environment information¸ and receives and actuates motion commands in the physical world.

Human-CAV Interaction (HCI) assumes that the CAV is impersonated by an avatar, whose model is selected/produced by the CAV owner or renter. The visible features of the avatar are head, face, and shoulders, and the audible feature is speech.

This use case requires the following functions:

1. Outside the CAV, when a group of humans approaches the CAV:
   1. The HCI creates AV Scene Descriptors providing access to the individual audio and visual objects, in particular the speech and the face of the humans from the rest of the Environment (close to the CAV).
   2. The HCI authenticates the humans it is interacting with by recognising the human CAV rights holder using Speech and Face Descriptors.
2. Inside the CAV, when a group of humans is sitting in the seats of the cabin:
   1. The HCI creates an AV scene description that separates and locates the visual and speech part of the humans from the rest of the Environment (cabin).
3. Two ways of human-HCI interaction in the cabin in two ways:
   1. By responding to commands and queries from one or more humans at the same time,
   2. By conversing with and responding to questions from one/more humans.
4. The HCI captures the humans’ Text and Personal Status.
5. The HCI shows itself as an avatar’s face and shoulder by means of a Personal Status Display.

Diagram

Description automatically generated with low confidence

Figure 8 – Human-CAV Interaction Reference Model

# Avatar-Based Videoconference

Avatar-Based Videoconference is a system composed of the 4 subsystems depicted in Figure 9 where participants entrust the image of the upper part of their bodies to avatars who utter the participants’ real voice.

Graphical user interface, chart

Description automatically generated

Figure 9 – Avatar-Based Videoconference

With reference to Figure 9, the elements characterising the Use Case are:

1. The Virtual Environment is equipped with a table and an appropriate number of chairs for avatars:
   1. Representing the upper part (waist up) of participating humans with high similarity.
   2. Not corresponding to any participant, in particular a Virtual Secretary.
2. Each participant uses their transmitting client to send to the server:
   1. The participant’s avatar model, language preference, and speech.
   2. The Avatar Descriptors of the participant.
3. The Server
   1. Authenticates all participants using Speech and Face Descriptors.
   2. Sends all participants the Virtual Environment Model.
   3. Forwards to all participants the avatar Models and Descriptors.
   4. Translates the active speech signals to the requested languages.
   5. Sends speech signals to participants according to their language preferences.
4. The Virtual Secretary
   1. Collects statements made by participating avatars using avatars’ Personal Statuses.
   2. Makes a Summary by combining all information.
   3. Makes the Summary available for avatars to read and edit the Summary.
   4. Edits the Summary based on Personal Status-based conversations with avatars.
5. Each participant uses their receiving Clients to:
   1. Place all avatar models on a chair of their choice.
   2. Animate the avatar models with the received Descriptors.
   3. Attach the speech signals to the mouths of the corresponding avatars.
   4. Select a Point of View, not necessarily the position assigned to his/her avatar.
   5. Watch the Virtual Environment and listen to the resulting spatial audio.

Figure 10, Figure 11, Figure 12, and Figure 13 represent the Reference Models of the Transmitting Client, the Server, the Personal Secretary, and the Receiving Client.

Diagram, text

Description automatically generated with medium confidence

Figure 10 – Reference Model of Avatar Videoconference Transmitting Client

Graphical user interface, diagram

Description automatically generated

Figure 11 – Reference Model of Avatar-Based Videoconference Server

Chart

Description automatically generated

Figure 12 – Reference Model of Virtual Secretary

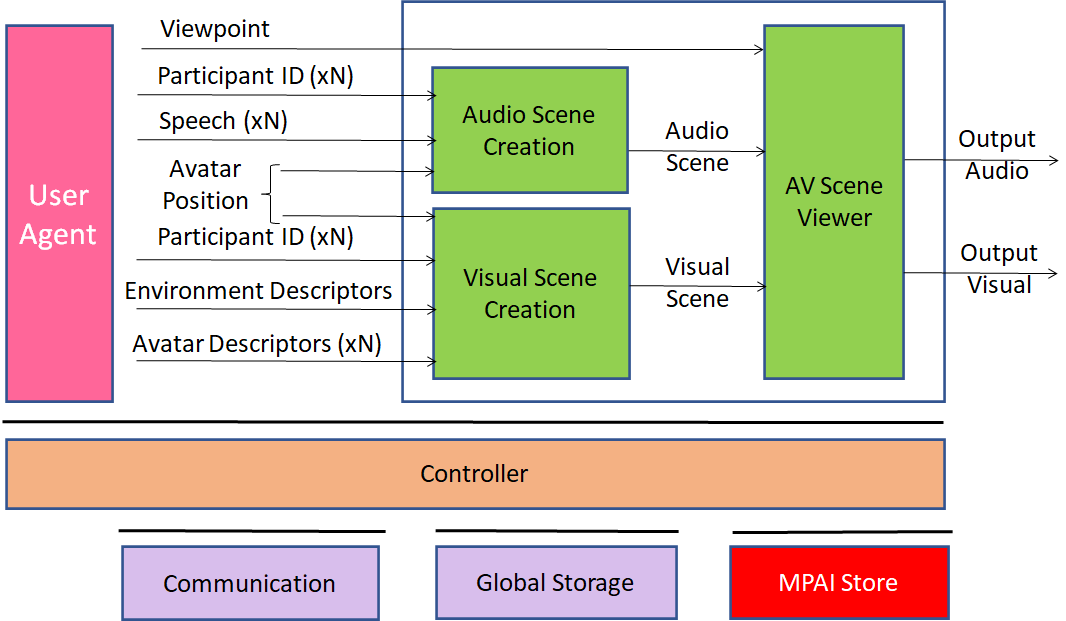


Figure 13 – Reference Model of Avatar-Based Videoconference Receiving Client